Explainable Structuring and Discovery of Relevant Cases for Exploration of High-Dimensional Data

Joris Falip
joris.falip@univ-reims.fr
CReSTIC
Reims, France

Frédéric Blanchard
CReSTIC
Reims, France

Michel Herbin
CReSTIC
Reims, France

ABSTRACT

Data described by numerous features create a challenge for domain experts as it is difficult to manipulate, explore and visualize them. With the increased number of features, a phenomenon called "curse of dimensionality" arises: sparsity increases and distance metrics are less relevant as most elements of the dataset become equidistant. The result is a loss of efficiency for traditional machine learning algorithms. Moreover, many state-of-the-art approaches act as black-boxes from a user point of view and are unable to provide explanations for their results. We propose an instance-based method to structure datasets around important elements called *exemplars*. The similarity measure used by our approach is less sensitive to high-dimensional spaces, and provides both explainable and interpretable results: important properties for decision-making tools such as recommender systems. The described algorithm relies on *exemplar theory* to provide a data exploration tool suited to the reasoning used by experts of various fields. We apply our method to synthetic as well as real-world datasets and compare the results to recommendations made using a nearest neighbor approach.

CCS CONCEPTS

• **Information systems** → **Decision support systems**; • **Human-centered computing** → *Interactive systems and tools*; • **Computing methodologies** → *Knowledge representation and reasoning*.

KEYWORDS

instance-based algorithm; high-dimensional data; exploratory analysis; information visualization; exemplar theory; explainable machine learning; recommendation system

ACM Reference Format:
Joris Falip, Frédéric Blanchard, and Michel Herbin. 2019. Explainable Structuring and Discovery of Relevant Cases for Exploration of High-Dimensional Data. In *Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, March 20, 2019* , 7 pages.

1 INTRODUCTION

This work addresses the task of structuring a dataset to facilitate exploration and visualization. The proposed idea is to create pairwise connections between similar elements in order to provide a similarity graph usable for exploration and recommendation. Our goal is a solution well-suited for domain experts, that takes into account and even leverages the analogical reasoning most experts use: they often compare new cases or problems to previously encountered ones and base their decisions on past experiences. This way, the tool and answers it provides would feel more intuitive for end-users. We also emphasize compatibility with high-dimensional data: experts often rely on datasets described by hundreds or thousands of features, meaning that any solution must remain efficient even with a high number of dimensions. Finally, explicability and interpretability are critical in many fields including health and medicine, insurance, finance or security: a recommendation system is even more useful when end-users understand the process behind suggested decisions. It allows them to adapt and adjust the algorithm to their needs but most of all, it increases confidence in the results.

To reach these objectives, we created an unsupervised instance-based algorithm that enriches data and structures it, linking each element to an exemplar exhibiting typical characteristics. The resulting graph can be used by a recommender system, building a solution to explore data by following edges, each one representing strong similarities on specific features. When exploring the dataset, going from an instance to its exemplar means going from a specific element to a more generic and archetypal one. Our approach processes each dimension individually then aggregate the results, making this method suitable for very high-dimensional analysis. Moreover, it identifies features contributing to similarities between elements, making it straightforward to explain the structuration choices. This strategy can be adapted to multiple contexts as it does not require any *a priori* knowledge.

In the following section, we detail the problems encountered when manipulating high-dimensional datasets and elaborate on the requirements of a successful exploration tool that would allow visualization of similarities between data points and could give insights on hidden patterns. The next section describes the proposed algorithm, the resulting structure and how the input parameter influences it. We then detail applications to various datasets, illustrating the benefits of this method. Finally, we discuss the proposed approach, review current work in progress and possibilities for future work.

2 BACKGROUND

While high-dimensional databases are frequently encountered with many real-world data, experts lack appropriate and simple tools to make the best use of these datasets and information. Many knowledge discovery tools provide relevant results but lack interpretability or explainability: essential features when making any critical decision [9].

IUI Workshops'19, March 20, 2019, Los Angeles, USA

Also, when exploring and visualizing data, most algorithms rely on *prototypes* [15] generated with averages from similar elements. These prototype-based methods can be effective in some cases but when manipulating data, experts of a specific field tend to understand a new element (or *instance*) by comparing it to the most similar instance they are familiar with [8, 16]. Experts naturally favor this approach when confronted with problems within their field of expertise, and it provides better results. [10]. This reasoning is known as analogical reasoning and is formalized by the *exemplar theory* [11, 12]. For example, a physician will use his work experience to diagnose a new patient by linking it to the most similar element in a set of typical patients he encountered in the past. In analogical reasoning, typical instances of a dataset are called *exemplars*, each one subsuming a set of similar elements. These exemplars form a subset that can describe the whole dataset but remains small enough to manipulate and explore easily.

When analyzing data described by a high number of features, it becomes harder to use the distance between instances as a measure of similarity. This is because working with data described by a high number of features comes with two inherent problems: data become sparser as density decreases and distances between elements normalize, so neighbors are harder to distinguish [1]. These phenomena, often referred to as "curse of dimensionality" [5], make many analysis and recommendation algorithms unreliable [4] because they use distances between instances to compute similarities. A standard solution to this problem is to use feature projection methods to reduce dimensionality, but this is not suitable if we wish to retain the interpretability of the process. On the other hand, feature selection reduces the genericity of the approach and require time and *a priori* knowledge to select and filter attributes.

High-dimensionality also provides new properties to elements, like the hubness phenomenon [18], which can help to improve and develop data mining techniques. Hubness is the fact that increased dimensionality correlates with some instances being in the nearest neighbors of the majority of the other elements, turning these instances into so-called hubs. Hubness reduction methods, when coupled with traditional approaches, do not improve their results [7]. Hopefully this phenomenon can be used to adapt algorithms to high-dimensional data [17].

3 STRUCTURING AROUND EXEMPLARS

Our proposed approach finds the most typical and "interesting" neighbor of each data point, without relying on a distance metric in the entire space [19]. We process each dimension separately before aggregating the results: our algorithm is less affected by high dimensionality, and interpretability is improved as we can quantify the importance of each dimension regarding associations. We also chose to use ranks instead of distances, to avoid excluding outliers elements.

Structuring a dataset by linking each element to its exemplar results in a graph where each connected component is a group of elements similar enough to be subsumed by a few selected exemplars. This graph allows for exploration of the dataset, with edges representing strong similarity between individuals.

From a practical point of view, the proposed method is currently implemented as a recommender system used by medical experts. This allows them to visualize and understand diabetes-related data by exploring an association graph [6] where each vertex is a patient and patients in the same connected component had a similar evolution of the disease.

Our algorithm is based on the *Degree of Representativeness* [2] (*DoR*). The *DoR* measures the typicality of an element and its ability to be an exemplar for other elements. *Algorithm 1* illustrates the selection of an exemplar for each element according to the *DoR*, while *Algorithm 2* details computing of the *DoR*. This new algorithm addresses the problem of high-dimensionality: our last work was using distance in the full description space when computing the *DoR*, thus being unreliable with datasets described by a high number of features.

3.1 Algorithm

Let Ω be a set of N objects in a multidimensional space. These N objects, or elements, are characterized by D qualitative or quantitative features.

For each of the D dimensions, we compute the distance matrix. Each object then ranks every other object according to their similarity on this dimension: low distance translates to a good ranking, with the nearest element being ranked 1 and the farthest ranked $N-1$. This step is repeated on each dimension to transform the D distance matrices into D rank matrices.

Let us transform the ranks into scores. Let x be an object of Ω: for each dimension d, x assigns a relative score $Score_x^d$ to every other object y of Ω. $Score_x^d$, relative to x, can be any arbitrary function, but in this paper it will be defined by:

$$Score_x^d(y) = \max(1 + T - Rank_x^d(y), 0) \tag{1}$$

where $Rank_x^d(y)$ is the rank of an object y relative to x on dimension d, and each element only assigns a non-zero score to its T nearest neighbors. For each element y, we can compute the sum of all scores it received on a specific dimension:

$$Score^d(y) = \sum_{x \in \Omega} Score_x^d(y) \tag{2}$$

. Let us define k as a numeric parameter allowing control over the total number of exemplars. To find the exemplar of a given object x, we introduce the $DoR_x(y)$ of another element y as the sum of the scores $Score^d(y)$ for every dimension d where y is in the k nearest neighbors of x.

$$DoR_x(y) = \sum_{d \in D'} Score^d(y) \tag{3}$$

where D' is the set of dimensions on which y is in the k-nearest neighbors of x. The element chosen as an exemplar for object x is the element with the highest DoR_x.

Data: N elements defined on D dimensions, neighborhood size K
Result: list of N exemplars
foreach *dimension d in D* **do**
  Compute dissimilarity matrix of elements;
  Transform similarities into ranks;
  Convert ranks into scores with arbitrary scoring function;
  foreach *element e in N* **do**
    $Score^d(e) \longleftarrow \sum_{\forall x \in N} Score_x^d(e)$ given by x to y;
  end
end
foreach *element e in N* **do**
  $exemplar(e) \longleftarrow \max_{\forall x \in N}(\mathbf{DoR}(x, e, K))$;
end
Result $\longleftarrow$ list of exemplars determined above;

Algorithm 1: Structuring algorithm

Data: elements x and e defined on D dimensions, neighborhood size K
Result: Degree of Representativeness of element x according to element e
$Knn^d(e) \longleftarrow$ *K-nearest neighbors of e, on dimension d;*
$DoR \longleftarrow 0$
foreach *dimension d in D* **do**
  if $x \in Knn^d(e)$ **then**
    $DoR \longleftarrow DoR + Score^d(x)$;
  end
end
$Result \longleftarrow DoR$;

Algorithm 2: *DoR* computing algorithm

3.2 Neighborhood size k and the resulting structure

When establishing pairwise connections between elements, parameter k determines the total number of exemplars used to summarize the dataset. A low value for the neighborhood size (k about 20% of the number of elements) gives numerous exemplars, each one closely matching the few instances they subsume. The high number of connected components in the resulting graph limits exploration of related instances by following edges but this configuration will only group very similar elements. On the other hand, selecting a higher value (about 30% to 50% of the dataset's size) gives fewer exemplars that each subsumes a significant subset of the population. These exemplars provide a meaningful but condensed representation of the data and exploration of the resulting structure is easy: with few connected components in the graph, following edges allows the discovery of many similar elements. *Figure 6*, discussed in the results section, illustrates the number of connected components obtained for every possible value of k on a real-world dataset and thus the different granularities obtainable in the final graph.

Figures 1 and *2* illustrate these differences: the former uses a neighborhood size of 30 and is composed of 44 exemplars, with an average

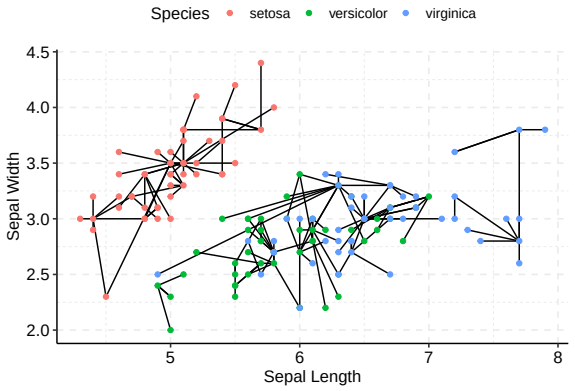

Figure 1: Application to the Iris dataset with neighborhood size k set to 30 create numerous exemplars closely matching elements they subsume.

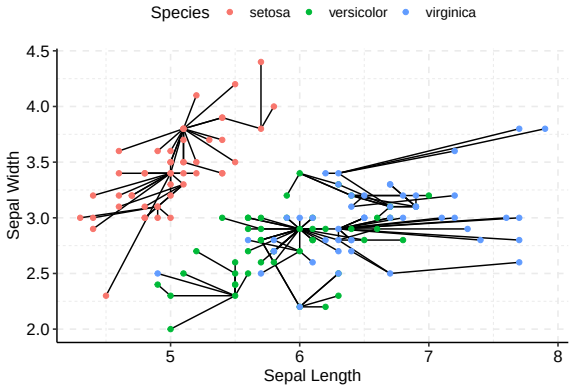

Figure 2: Application to the Iris dataset with neighborhood size k set to 50 create few exemplars that summarize the entire dataset.

of 2 elements subsumed by each exemplar; while the later is obtained with a parameter set to 50, resulting in 25 exemplars for an average of 5 elements represented by each exemplar. Either one of these two graphs makes for an excellent recommendation system where, for each element, a similar but more generic and typical one can be suggested. It is also useful as a visualization tool, guiding users as they explore the dataset. Features playing the most prominent role in each association can even be overlaid on the edges to inform on the nature of the similarities guiding the exploration process.

Given the graph resulting from the structuring process, another option is to study its evolution for every value of the parameter k. For a set of N elements, this means $1 \leq k \leq N$. From this, we can also establish other measures including the total number of iterations a given element is chosen as an exemplar, or the average number of elements subsumed by each exemplar. These additional

insights put more perspective on the usefulness of each instance as an exemplar and its ability to subsume other elements.

4 EXPERIMENTS

To evaluate the efficiency of the described approach, we analyze graphs created by our structuration algorithm and compare them to graphs generated using the nearest neighbor algorithm in high-dimensional datasets. We compare both methods using key indicators related to connected components in the graphs, to study the usefulness of each structure as a recommendation tool in a data exploration scenario. *Table 1* summarizes the studied datasets along with their respective number of elements and dimensions.

dataset	topic	elements	dimensions
normal	N/A	300	1000
residential [13]	real estate	372	103
communities [14]	law enforcement	2 215	101

Table 1: Description of the synthetic and real-world datasets used for comparison of the graph structure given by the proposed algorithm and a nearest neighbor selection.

normal is a synthetic dataset composed of 300 elements described on 1000 dimensions. 80% of their dimensions are sampled from a normal distribution centered on 0 with a standard deviation of 1. The remaining 20% dimensions are noise sampled from a uniform distribution between 0 and 1. This very high-dimensional example outlines the previously mentioned "curse of dimensionality": with a thousand dimensions, it is a sparse dataset and distances between elements become hard to distinguish.

Residential Buildings [13] (abbreviated "*residential*") and *Communities and Crime* [14] (abbreviated "*communities*") both are datasets made available through the *UCI* machine learning repository [3]. They are described with more than a hundred features, some of them being potential goals for prediction algorithms. *residential* contains physical, financial and economic data related to 372 construction projects of residential apartments in Tehran, the capital and largest city of Iran. *communities* combines data related to law enforcement as well as crime and its socio-economic factors. It aggregates information about 2215 cities from the United States, collected between 1990 and 1995.

Before any analysis, we removed prediction goals and irrelevant features (such as elements IDs) from both datasets to guarantee that two similar elements will be close to one another in the full description space. Dimensions containing missing values were also pruned in the preprocessing step, representing a total of 2 and 46 features removed from the *residential* and *communities* datasets respectively, the later containing numerous features that are prediction goals. This was done to avoid noise that would be detrimental mostly to the nearest neighbor method. The number of dimensions noted in *Table 1* accounts for the remaining dimensions.

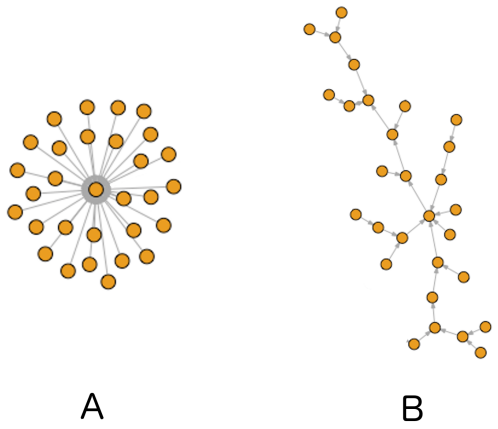

Figure 3: Two connected components featuring the same number of elements. Structuration A) does not provide insights regarding similarities, while B) allows step by step exploration with growing typicality when starting from any element.

To create the similarity graph with our structuration around exemplars, we set to 10 the parameter T from *Formula 1*, meaning the scoring function will give a positive score to the ten nearest neighbors of each instance. As a reminder, this function is chosen arbitrarily as a generic scoring function but can be adapted for a specific dataset, given some *a priori* knowledge. The parameter we can fine-tune is the neighborhood size used in the last step of our algorithm when choosing an exemplar for each element. However, because this parameter allows for a tradeoff between edges representing very strong similarities and graphs exhibiting structure and properties suitable for exploration (low number of connected components with high diameter), we will not seek to optimize this parameter to avoid skewing the comparison. As suggested in the previous section, we define k as 30% of the number of elements in the dataset so we can expect the right balance between meaningful associations and optimal graph structure.

This graph is compared to another graph generated by linking each element of the dataset to its nearest neighbor using the Minkowski distance of order 0.75. While using an order lower than 1 violates the triangular inequality, it provides significantly better results in high-dimensional spaces than Euclidian or even Manhattan distances [1]. This way we can be sure that elements linked together by the nearest neighbor method will be as close as possible in the full dimensional space.

We select three criteria to evaluate the suitability of each graph as an exploration tool. The first two are the number of connected components and their size: they inform us on whether, starting from an element, following edges representing similarities will allow visualization of a broad selection of related instances or if exploration will quickly stop due to a lack of similar elements. We also study the average diameter of the components for each graph: a lower diameter indicates that most elements chose the same exemplar, as in *Figure 3.a.* A greater diameter, similar to the structure

	exemplar structuring			nearest neighbor		
dataset	number	size	diameter	number	size	diameter
normal	18	16.7	5.2	22	13.6	2.8
residential	64	5.8	2.7	105	3.5	2.1
communities	256	8.7	3.2	400	5.5	2.5

Table 2: Analysis of graphs obtained with the proposed approach and a nearest neighbor selection approach. The number of connected components and their mean size and mean diameter are detailed.

illustrated in *Figure 3.b*, gives the possibility to progressively explore from a specific element to the most typical one.

5 RESULTS

Results of our experiments are shown in *Table 2*. For each method and dataset we display the number of connected components obtained, compute their mean size and their mean diameter. We can see that, even for very high-dimensional data, our approach results in better structuring of the instances where elements are linked to another more archetypal exemplar within each component. This is outlined by the smaller number of connected components and their overall larger diameter. This trend can be seen for every dataset we studied: reducing the number of components by 18% to 39% and demonstrating promising results on the two real-world applications.

Figure 4 shows the structure obtained with the nearest neighbor approach on the *residential* dataset. In this case, data are weakly structured, making it hard for a user to gain any insight or to use this graph for exploration. *Figure 5* illustrates the same dataset where instances are structured around exemplars. The structure makes more sense if used to visualize similarities: connected components contain more elements and have a larger diameter. Navigating data is easier and more instances are available as exemplars so users can choose the level of genericity they require for an exemplar, by following multiple edges to an exemplar subsuming more nodes. Such an exemplar will be farther from the initial instance, but more archetypal and may better suit analogical reasoning by exhibiting typical characteristics.

With *Figure 6* we illustrate the number of connected components in the graph created with various values for the parameter k. This figure demonstrates how k can be used as a granularity factor to choose on a spectrum ranging a few components subsumed by very archetypal exemplars to numerous subsets composed of closely matching elements. The maximum number of components for any value of k is equal to the number of components created by the nearest neighbor approach, meaning that even for a very low of k the proposed structure is no less suitable for exploration than the nearest neighbor one.

To provide insights on how the data is structured, we detail edges from a small connected component shown in *Figure 7*. This component was extracted from the *residential* dataset, with the same parameters previously used for *Figure 5*: neighborhood size k of 124 and the threshold T from *Formula 1* set to 10. We can describe the following:

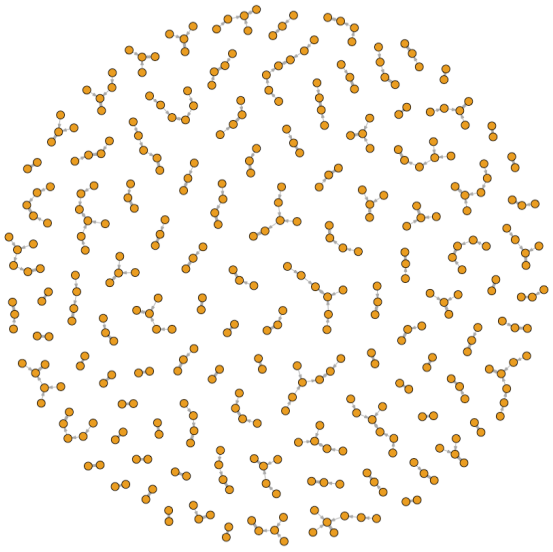

Figure 4: Structuring *residential* data by linking every element to its nearest neighbor using a Minkowski distance of order 0.75. There are numerous connected components with only two or three elements: these are not suitable for exploration based on similarity.

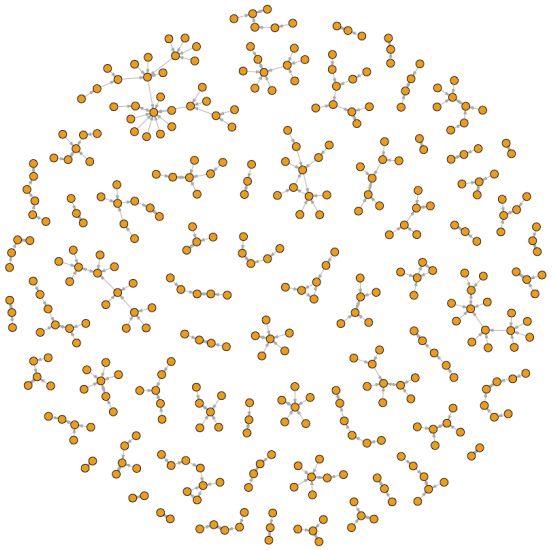

Figure 5: Structuring *residential* data using exemplars to create meaningful associations. This creates bigger connected components containing similar instances.

- Vertices **1** and **2** both chose vertex **3** as their exemplar because **3** is in the k-nearest neighbors of those elements on 89 dimensions, so they each are similar to **3** regarding most

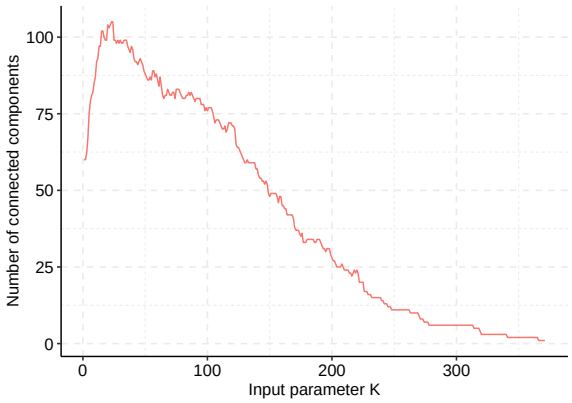

Figure 6: Evolution of the number of connected components depending on the input parameter k, when structuring the *residential* dataset.

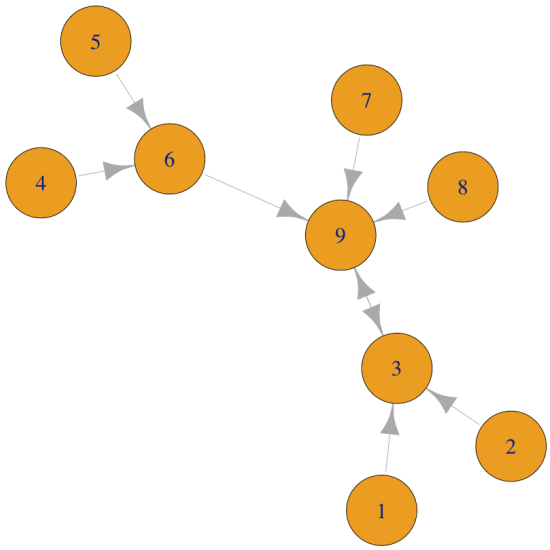

Figure 7: A connected component extracted from the result shown in *Figure 5*. The labels are arbitrary and added for clarity.

features of the dataset. The dimensions linking **1** to **3** are the same as those linking **2** to **3**.

- Elements **1** and **2** are similar to each other on 96 features. However, they have not chosen each other as exemplars because they both gave instance **3** a cumulated score of more than 4,600 while they evaluated each other at only around 2,000.
- **3** is similar to **9** on 96 features, attributing it a total score of 5,000. As **3** only had a score of 4,600, **9**'s higher score confirms that it is indeed more typical. **1** and **2** were similar to **3** on 89 features, from which 88 are also among the 96 similar features between **3** and **9**.

6 CONCLUSION

To summarize this work, we proposed an instance-based structuration method that computes similarities between complex elements. It creates pairwise connections between related instances with exemplars subsuming similar elements. This structure allows for intuitive exploration of a dataset and is relevant when creating a recommendation system. We validated our approach by studying a simulated dataset and data from real estate market and law enforcement. We compared our results to a nearest neighbor approach suited for high-dimensionality where proximities in the full dimensional space were computed using a Minkowski distance of order 0.75 to avoid concentration of the distances. On these high-dimensional examples we outlined better performances that are less subject to the "curse of dimensionality" usually encountered with these types of data. We also experimented on the use of parameter k as a granularity factor giving users a straightforward way to influence similarities and the structuring process.

We are currently testing this algorithm's results when applied to medical data, to help endocrinologists to better understand diabetes and diagnose patients more accurately. Future works include further development of our currently deployed prototype to easily gather user feedback on each association. This feedback could be used to create an automatic weighing of the features, tailoring the recommendation to each user's preferences.

6.1 Source code

The source code implementing the proposed algorithm and used to create the figures and results in this paper is available on the following public Github repository: https://github.com/Elesday/ESIDA-IUI19.

REFERENCES

[1] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. 2001. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In *Database Theory — ICDT 2001*. Springer Berlin Heidelberg, Berlin, Heidelberg, 420–434.
[2] Frédéric Blanchard, Amine Aït-Younes, and Michel Herbin. 2015. Linking Data According to Their Degree of Representativeness (DoR). *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems* 2, 4 (June 2015), e2.
[3] Dua Dheeru and Efi Karra Taniskidou. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml
[4] Pedro Domingos. 2012. A few useful things to know about machine learning. *Commun. ACM* 55, 10 (2012), 78–87.
[5] David L Donoho et al. 2000. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture* 1 (2000), 32.
[6] Joris Falip, Amine Aït Younes, Frédéric Blanchard, Brigitte Delemer, Alpha Diallo, and Michel Herbin. 2017. Visual instance-based recommendation system for medical data mining.. In *KES*. 1747–1754.
[7] Roman Feldbauer and Arthur Flexer. 2018. A comprehensive empirical comparison of hubness reduction in high-dimensional spaces. *Knowledge and Information Systems* (May 2018), 1–30.
[8] Gary Klein, Roberta Calderwood, and Anne Clinton-Cirocco. 2010. Rapid Decision Making on the Fire Ground: The Original Study Plus a Postscript. *Journal of Cognitive Engineering and Decision Making* 4, 3 (Sept. 2010), 186–209.
[9] Zachary C Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490* (2016).
[10] Michael L Mack, Alison R Preston, and Bradley C Love. 2013. Decoding the Brain's Algorithm for Categorization from Its Neural Implementation. *Current Biology* 23, 20 (Oct. 2013), 2023–2027.
[11] Douglas L Medin and Marguerite M Schaffer. 1978. Context theory of classification learning. *Psychological Review* 85, 3 (1978), 207–238.
[12] Robert M Nosofsky. 1986. Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General* 115, 1 (1986), 39–57.

[13] Mohammad Hossein Rafiei and Hojjat Adeli. 2015. A novel machine learning model for estimation of sale prices of real estate units. *Journal of Construction Engineering and Management* 142, 2 (2015).
[14] Michael Redmond. 2009. Communities and crime data set. *UCI Machine Learning Repository* (2009).
[15] Eleanor H Rosch. 1973. Natural categories. *Cognitive psychology* 4, 3 (1973), 328–350.
[16] Jeffrey N Rouder and Roger Ratcliff. 2006. Comparing Exemplar- and Rule-Based Theories of Categorization. *Current Directions in Psychological Science* 15, 1 (2006), 9–13.
[17] Nenad Tomasev and Dunja Mladenic. 2011. Nearest Neighbor Voting in High-Dimensional Data: Learning from Past Occurrences. In *International Conference on Data Mining Workshops*. IEEE, 1215–1218.
[18] Nenad Tomasev, Milos Radovanovic, Dunja Mladenic, and Mirjana Ivanovic. 2014. The Role of Hubness in Clustering High-Dimensional Data. *IEEE Transactions on Knowledge and Data Engineering* 26, 3 (2014), 739–751.
[19] Yingzhen Yang, Xinqi Chu, Feng Liang, and Thomas S Huang. 2012. Pairwise Exemplar Clustering. *AAAI* (2012).